Klausur (Bio-Informatik)

Einführung in die Datenbanksysteme
Datenbanken für die Bioinformatik

Heinz Schweppe, Joos-Hendrik Böse, Manuel Scholz

1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*Σ*
25	**20**	**16**	**12**	**11**	**04**	**07**	**07**	**102**

Name:_____

Matrikelnummer:_____

Zu beachten:

- Namen und Matrikelnummer eintragen.
- Es sind keine Unterlagen erlaubt.
- Bitte sämtliche Lösungen auf die Blätter (auch Rückseite) der Klausur schreiben. Eigene Blätter sind nicht erlaubt.
- Täuschungsversuche jeglicher Art führen zum Nichtbestehen der Klausur (0 Punkte)

1. Aufgabe *(2+1+4+1+3+1+4+2+3+4 Punkte)*

(Anmerkung: Bei den Multiple Choice Fragen der Klausur sind zum Teil auch mehrere Antworten möglich. Für jede richtig angekreuzte Aussage gibt es Punkte, für falsch angekreuzte Aussagen Punktabzug.)

- Welchen Unterschied gibt es zwischen Primär - und Sekundärindex (ein Satz)

Primärindex: physikalische Anordnung der Elemente
Sekundärindex: logische Anordnung der Elemente (auf dem Primärindex)

- Wann kann es mehr als einen Primärindex geben?
  ☐ nie ☐ wenn es mehr als einen Kandidatenschlüssel gibt ☐ immer *nie*

- Nennen Sie je zwei Vor- und Nachteile der Raid-Plattenorganisation (Level 4: Blockverschachteltung - block striping - , eine Paritätsplatte):

1. Vorteil: *Durchsatz beim Lesen wird erhöht*

2. Vorteil: *Sicherheit, Recovery beim Ausfall einer Platte*

1. Nachteil: *Jeder Schreibvorgang benötigt 2 Schreiboperationen (normal + parity disk)*

2. Nachteil: *Parity Disk ist Hotspot*

- Wie viel physische Schreiboperationen erfordert eine logische Schreiboperation (Schreiben eines Datums) bei Raid-level 5 (Paritätsblöcke verteilt auf Daten-Laufwerke)?

  Zwei, einen auf der normalen Platte und einen auf der zuständigen Parity Platte.

- Eine SQL Anfrage besteht aus maximal 6 Klauseln (weitgehend unabhängigen Bestandteilen des Befehls), welche sind das?

  SELECT, FROM, WHERE, GROUP BY, HAVING, ORDER BY

- Aus wie vielen Klauseln muss eine SQL Anfrage mindestens bestehen?

  Zwei, und zwar: SELECT, FROM

- Selektion, Kreuzprodukt, Projektion, Mengendifferenz und Vereinigung bilden eine Basis der relationalen Algebra (d.h. alle anderen Operationen der RA lassen sich durch diese ausdrücken). Ersetzen Sie ein Element der Basis, so dass wiederum eine Basis entsteht und begründen Sie in einem Satz (oder einer RA-Formel) warum es sich um eine Basis handelt.

  Anstatt Kreuzprodukt einfach Join mit Prädikat true.

- Was versteht man unter referentieller Integrität? (Ein Satz)

  Verwendung eines Schlüssels einer Relation als Attribut einer anderen Relation. (Fremdschlüssel, FOREIGN KEY)

- Geben Sie ein möglichst kleines Fragment von Definitionen (Create Table) von zwei Tabellen A und B in SQL an, in der referentielle Integrität spezifiziert wird.

  CREATE TABLE A (x NUMBER PRIMARY KEY);
  CREATE TABLE B (y NUMBER FOREIGN KEY REFERENCES A(x));

- Können bei der wechselseitig referentiellen Integrität zwischen zwei Tabellen A und B Probleme auftreten? Begründen Sie je nach Antwort *warum nicht* bzw. *wie das Problem behoben werden kann.*

  Theoretisch können Probleme auftreten, weil beides gleichzeitig eingefügt werden müsste (Chicken – Egg – Problem). Mit Deferred kann das jedoch umgangen werden.

2. Aufgabe *(9X1,5 [Ent,Rel] + 2x2,5 [Weak, General] +0,5 Punkte)*

- Betrachten Sie folgende Beschreibung eines Tierheims:
  In dem Tierheim gibt es Käfige, die eine eindeutige Nummer und eine Größenangabe in m² besitzen. Pro Käfig wird immer nur ein Tier gehalten, d.h. es darf pro Tag für einen Käfig nicht mehr als ein Tier zugewiesen sein. Für jedes Tier gibt es eine eindeutige Nummer, eine Gewichtsangabe und einen Namen. Es wird genau festgehalten, welches Tier an welchem Tag einen Käfig belegt hat, da die Tiere oft verlegt werden.
  Natürlich gibt es in dem Tierheim auch Mitarbeiter, die durch ihre Sozialversicherungsnummer eindeutig zu bestimmen sind. Für jeden Mitarbeiter wird zusätzlich das Gehalt festgehalten. Es gibt einerseits Pfleger, die die Käfige reinigen und für die bekannt ist, ob sie eine Tierpflegerausbildung abgeschlossen haben oder nicht. Jeder Pfleger hat mehrere Käfige für deren Reinigung nur er zuständig ist. Weiterhin gibt es Verwaltungspersonal, für das die Gesamtnote ihres letzten Zeugnisses festgehalten werden soll und ob es einen Tippkurs belegt hat oder nicht.

- Entwerfen Sie ein UMLER-Diagramm für den beschriebenen Sachverhalt. Identifizieren Sie alle notwendigen Entitäten und Beziehungen. Tragen Sie die Kardinalitäten in Min-Max Notation ein. Kennzeichnen Sie die Schlüssel der Entitäten. Fügen Sie keine künstlichen Schlüssel ein.
  Benutzen sie für den Entwurf des UMLER-Diagramms die letzte Seite, welche auch abgetrennt werden kann. *Achten Sie darauf, dass die Seite mit ihrer Matrikelnummer und ihrem Namen beschriftet ist!*

3. Aufgabe *(3+3+4+3+3 Punkte)*

Gegeben sind die folgenden Relationen:

A (id, datum, schlagzeile, autor, text)

SW (id, schlagwort)

Inhalt der Datenbank sind Artikel, z.B. die einer Nachrichtenagentur. (Relation A). Die Relation SW enthält Schlagwörter und ihr Vorkommen in den Artikeln.

- Formulieren Sie in SQL:

  Alle Schlagzeilen von Artikeln, in denen das Schlagwort „Irak“ vorkommt

```
SELECT schlagzeile
FROM A, SW
WHERE A.id = SW.id
AND SW.schlagwort = 'Irak';
```

  Schlagzeile und Autor für alle Artikel mit weniger als 10 Schlagwörtern

```
SELECT schlagzeile, autor
FROM A, SW
WHERE A.id = SW.id
GROUP BY schlagzeile, autor
HAVING COUNT(*) < 10;
```

- Formulieren Sie in relationaler Algebra (Es darf zusätzlich zu den Grundoperationen π, σ, X, -, U auch die Joinoperation verwendet werden):

  Die Namen der Autoren, die nie einen Artikel mit dem Schlagwort „Saddam“ geschrieben haben.

$$\Pi_{autor}(A) - \Pi_{autor}(\sigma_{schlagwort='Saddam'}(A \underset{P}{\bowtie} SW))$$

where $P = A.id = P.id$

- Wenn der Wert des Attributs „text“ NULL ist, soll der Wert des Attributs „schlagzeile“ auch Wert von „text“ sein. Formulieren Sie eine entsprechende Update-Operation.

```
UPDATE A set text = schlagzeile
WHERE text is null;
```

- Das Schema soll so erweitert werden, dass zu jedem „schlagwort“ die Häufigkeit des Vorkommens im „text“ und die Häufigkeit des Vorkommens in der „schlagzeile“ für jeden Artikel dargestellt werden kann. Ferner die Angabe, in wie viel Artikeln ein Schlagwort vorkommt. Geben Sie das erweiterte Schema an (Wie in der Aufgabenstellung, es muss nicht ausführlich mit einer CREATE TABLE Anweisung durchgeführt werden)

```
A (id, datum, schlagzeile, autor, text)

SW (id, schlagwort, anz_sw_text, anz_sw_schlagzeile,
     anz_artikel)
```

4. Aufgabe *(4+4+4 Punkte)*

Vergleichen Sie in den folgenden Aufgaben die Ergebnisse der Anfragen. Die Mengen können gleich, ineinander enthalten oder verschieden (nicht notwendig disjunkt) sein. Auf die Reihenfolge kommt es nicht an, wohl aber auf ggf. vorkommende Duplikate.

A)

Schema: R(x)

Q1:

```
SELECT x
FROM R rr
WHERE NOT EXISTS (SELECT x FROM R WHERE x > rr.x);
```

Q2:

```
SELECT MAX(x) FROM R;
```

☐ Die Antworten sind immer gleich.

☐ Die Antwort von Q1 ist immer in der Antwort von Q2 enthalten.

☐ Die Antwort von Q2 ist immer in der Antwort von Q1 enthalten. *XXX*

☐ Die Antworten sind immer verschieden.

B)

`Schema R(a,b)`, keine NULL-Werte in R aber evtl. Duplikate

Q1:

```
SELECT DISTINCT COUNT(*)
FROM R
GROUP BY a;
```

Q2:

```
SELECT DISTINCT COUNT(b)
FROM R
GROUP BY a;
```

☐ Die Antworten sind immer gleich. *XXX*

☐ Die Antwort von Q1 ist immer in der Antwort von Q2 enthalten.

☐ Die Antwort von Q2 ist immer in der Antwort von Q1 enthalten.

☐ Die Antworten sind immer verschieden.

C)

Betrachten Sie die Folge von Operationen Q1 bzw. Q2 auf der DB mit Schema R(a,b). Der Wert von Qi ist die Relation nach Ausführung von Q1 bzw. Q2.

Q1:

```
UPDATE R SET b = 3 WHERE b = 2;
```

Q2:

```
INSERT INTO R
SELECT a, 3 FROM R WHERE b=2;
DELETE FROM R WHERE b = 2;
```

☐ Q1 und Q2 sind gleich. *XXX*

☐ Q1 ist in Q2 enthalten.

☐ Q2 ist in Q1 enthalten.

☐ Q1 sind verschieden.

5. Aufgabe *(6+5 Punkte)*

A)

Gegeben seien die Relationen

R

A	B
0	1
4	5
8	9

S

B	C
1	2
5	2
5	6
5	10
13	10

T

C	D
2	3
6	7
10	11
10	3

- Wie viel Tupel liefert der *natürliche Verbund (join)* von **R**, **S** und **T**

  ☐ 5 *XXX* ☐ 8 ☐ 10 ☐ 13

- Wie viel Tupel liefert der *volle, natürlich äußere Verbund (full natural outer join)* zwischen **R** und **S** und dann **T**?

  ☐ 5 ☐ 8 *XXX* ☐ 13 ☐ 60

B)

Die Transaktionen T1 und T2 führen auf R(a,b) = {(2,3)} folgende Operationen durch:

T1:
```
INSERT INTO R VALUES (0,1)
DELETE FROM R WHERE a = 2 and b = 3;
```

T2:
```
SELECT * FROM R WHERE a >= 0
SELECT * FROM R WHERE b >= 0
```

- Die Transaktionen laufen mit dem Isolationslevel “repeatable read”. In welcher Reihenfolge die Operationen durchgeführt werden, ist nicht bekannt. Welche der folgenden Antworten auf **die beiden Anfragen von T2** sind *nicht* möglich? (Hinweis: Nicht an Sperrprotokolle o.ä. denken.)

☐ Erste Anfrage: {(2,3)}, zweite Anfrage {(0,1), (2,3)}

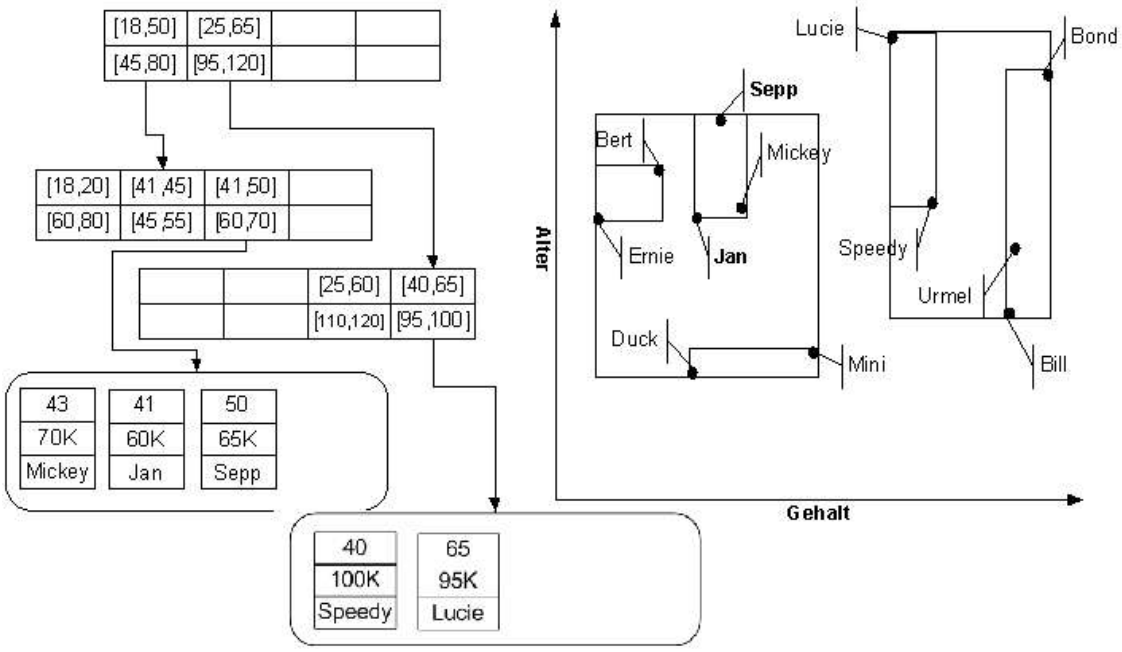
☐ Erste Anfrage: ∅ , zweite Anfrage {(0,1)} *XXX*

☐ Erste Anfrage: {(2,3)}, zweite Anfrage { (2,3)}

☐ Erste Anfrage: {(2,3)}, zweite Anfrage {(0,1)} *XXX*

6. Aufgabe *(4 Punkte)*

Gegeben sei der folgende R-Baum (links) und seine grafische Darstellung (rechts):



- Wieviel Blöcke müssen in dem gegebenen R-Baum gelesen werden, wenn folgende Bereichsanfrage gestellt wird:

```
SELECT * FROM AlterGehalt
WHERE Alter BETWEEN 43 AND 67
AND   Gehalt BETWEEN 65 AND 98;
```

Anzahl: *5*

Antwortmenge: *Sepp,Lucie, Mickey*

7. Aufgabe *(2+2+3 Punkte)*

A) Zeichnen Sie zu folgendem XML-Dokument den abstrakten Strukturbaum. Dieser enthält die Elemente des Dokuments und veranschaulicht die Schachtelung. Attribute werden als Annotation an die entsprechenden Dokumentknoten geschrieben. Die Werte der Elemente bzw. Attribute kommen im Strukturbaum nicht vor.

```
<?xml version="1.0"?>
<!--!DOCTYPE cdcollection SYSTEM "cd-collection.dtd"-->
<cdcollection>
  <album id="540 590-2">
          <title>Sheryl Crow</title>
          <artist>Sheryl Crow</artist>
          <label>A and M Records</label>
          <track time="4:56">maybe angels</track>
          <track time="3:50">a change</track>
          <track time="4:51">home</track>
          <track time="3:58">sweet rosalyn</track>
          <track time="5:23">if it makes you happy</track>
          <track time="4:27">redemption day</track>
          <track time="3:07">hard to make a stand</track>
          <track time="4:16">everyday is a winding road</track>
          <track time="4:43">love is a good thing</track>
          <track time="3:30">
            <composer> Lennon </composer> oh marie </track>
          <track time="4:58">superstar</track>
          <track time="4:34">the book</track>
          <track time="3:55">ordinary morning</track>
          <track time="3:20">free man</track>
  </album>
  <album id="332 80-2">
          <title>Slide on This</title>
          <artist realName = "Frank Peanut">Ronnie Wood</artist>
          <label>KOCH International</label>
          <track>Somebody Else Might</track>
          <track>Testify</track>
          <track>Ain't Rock'n Roll</track>
          <track>
            <composer> Gershwin </composer> Josephine</track>
          <track>Knock Yer Teeth Out</track>
          <track>Ragtime Annie (Lillie's Bordello)</track>
          <track>Must Be Love</track>
          <track>Fear For Your Future</track>
          <track>Show Me</track>
          <track>Always Wanted More</track>
          <track>Thinkin'</track>
          <track>Like It</track>
          <track>Breath On Me</track>
          <track>Somebody Else Might (Remix)</track>
  </album>
</cdcollection>
```

Strukturbaum:

Zu B): Syntax im Zusammenhang mit dem Schachteln von Elementen:

,	Aufzählung
|	Alternative
+	1 oder mehrere
*	0 oder mehrere
?	0 oder 1
nichts:	genau einmal

B) Ergänzen Sie für das obige Dokument die fehlenden Element- und Attributdefinitionen im folgenden DTD-Fragment für track, album und title. *(Hinweis: Es gibt mehrere Möglichkeiten. Einzige Bedingung: Konformität zu obigem Dokument)*

```
<!ELEMENT cdcollection (album*)>
<!ELEMENT artist (#PCDATA)>
<!ATTLIST artist realName CDATA #IMPLIED>
<!ATTLIST album id ID #REQUIRED>

<!ELEMENT album ......                    >
<!ELEMENT track (#PCDATA | composer?) >

<!ELEMENT label ......
```

C) Geben Sie ein Relationales Schema zur DTD aus B) an. Es muss nicht ordnungserhaltend sein.

8. Aufgabe *(4+3 Punkte)*

- Geben Sie mindestens vier Probleme an, die im Zusammenhang mit der Verwaltung von molekularbiologischen Daten auftreten und die sich in "klassischen" Datenbankanwendungen ("Kontenverwaltung") nicht oder nicht in gleichem Maße stellen. Erläutern Sie die Probleme ggf. in maximal zwei Sätzen.

(1)

(2)

(3)

(4)

- Anfragen im Vektorraummodell des Information Retrieval sind Mengen von Termen. Im Booleschen Modell Boolesche Ausdrücke von Termen. Erklären Sie warum.

Name: _____ Matrikelnummer: _____